REVISED
HANDBOOK FOR STUDYING THE EFFECTS OF THE LSC ON STUDENTS

Horizon Research
and
Westat

March 2001

TABLE OF CONTENTS

List of Appendixes

REVISED
HANDBOOK FOR STUDYING THE EFFECTS OF THE LSC ON STUDENTS

I. Purpose

The purpose of this handbook is to help Local Systemic Change (LSC) projects assess the effects of their activities on students and student learning. Recognizing the difficulty of measuring student impacts, these guidelines have been developed to help projects design studies that will meet both their own information needs and those of the National Science Foundation (NSF). The handbook addresses a number of important issues for research and evaluation studies, including deciding on appropriate measures, study design, data analysis, and reporting, with a particular emphasis on being able to make the case that any gains you may detect are attributable to the LSC. Appendix A provides a quick reference guide to the concepts included in this document. Appendix B contains a glossary of terms used in this document. Additional resources are listed in Appendix C.

II. Why Study Student Outcomes?

The goal of NSF's Teacher Enhancement program is "the improvement of science, mathematics and technology teaching and learning at pre K-12 grade levels." Projects funded under this program address improving teaching with the goal of enhancing student learning. Until recently, the LSCs have been assessed primarily in terms of their effects on teaching. Now that the program is becoming more mature, it is appropriate to take the next step and examine effects on students.

Stakeholders at all levels want to know about student outcomes. Evidence of student impact is important at the local level, where parents and the community pay close attention to how well our public schools are meeting their young peoples' needs; at the state level, where decisionmakers want to know how well systems are operating; and at the federal level, where policymakers closely monitor the nation's accomplishments. At all levels, building support for reform efforts rests strongly on showing that investments pay off in improving what students know and can do.

In line with this increased focus on student outcomes, federal agencies have placed greater emphasis on collecting sound data on student learning. NSF has identified student impacts as one of the critical indicators of the success of its programs. Specifically, one of the outcomes that it must demonstrate to Congress is "improved achievement in mathematics and science skills needed by all Americans." In order to do this, NSF looks to the projects it funds to provide such evidence.

It is important to note that, in looking at the effects of the LSC, one should interpret "effect on students" fairly broadly. In addition to examining student achievement, you may want to consider looking at the effect of the LSC on other student outcomes, such as student participation in higher level mathematics and science courses, attendance patterns, and attitudes towards mathematics and science. The examination of multiple outcomes in your studies, as well as the use of both quantitative and qualitative data collection methodologies are encouraged.

III. Attribution – Making the Case for Your Results

The LSC program is designed to improve the teaching and learning of science, mathematics, and technology by focusing on the professional development of teachers within whole schools or school districts. Projects are expected to designate the instructional materials to be used and then to provide extensive professional development to help teachers deepen their subject matter knowledge and become skilled in the use of the instructional strategies called for in those materials; they also are expected to provide support for teachers as they implement the instructional materials in their classrooms.

While all of the LSC projects share those elements of program design, projects were encouraged to develop intervention strategies that fit the needs of their particular target population and their particular context. Thus projects vary, for example, in the relative emphasis they give to teacher content knowledge, how they distribute the required hours of professional development over the course of the project, and the extent to which they provide professional development district-wide as opposed to at the school site.

Individual projects are being asked to assess the effectiveness of their strategies for students. Taken together, the results of these individual studies will provide valuable information on how effective the overall program strategy has been. When teachers are provided extensive professional development around the use of high quality instructional materials, do their students learn more? If the LSC does not lead to improved student learning, it will be difficult to make the case that the program should be continued. If, on the other hand, different researchers—studying variations of the LSC design in diverse contexts, using a variety of outcome measures—demonstrate the effect of the LSC on students, there will be good reason to continue and even expand the program.

In assessing the effect of the LSC on students, a key question that projects need to address is: "How do you determine with reasonable probability that the LSC, and not some other policy, program, or event was responsible for any student gains?" These guidelines are intended to help projects make the case that any growth they identify is, in fact, due to the LSC. Three conditions are needed to make a case for causality: temporal precedence, a correlation between treatment and outcomes, and a lack of plausible, alternative explanations.

First, it must be clear that the treatment occurred before the observed effect (i.e., "temporal precedence"). While this may seem obvious in most educational research, there are times when the order of events must be considered. When cyclic fluctuations occur, as often happens in economics, establishing a causal relationship can be difficult. In the case of the LSCs, you know when the professional development began, and you will have a measure of outcomes at some point after that, so the condition of temporal precedence is easily met.

Next, you have to show that there is a relationship between the treatment, professional development, and the effect, e.g., student scores, or participation in advanced mathematics/science courses, or some other outcome of interest. This relationship can be demonstrated by showing that if the program is provided, you have a particular outcome, and if the program is not provided, you don't. Perhaps more applicable to the LSC program, where projects work with all teachers, you need to show that providing more of the program leads to more of the outcome, while less of the treatment leads to less of an outcome. It should be emphasized that showing a relationship between the treatment—professional development training—and the outcome—student scores—is not sufficient to show that the treatment caused the outcome.

In order to support the likelihood of a causal relationship, you must rule out other possible explanations for the effect. Here is where the research design comes in, which is the central focus of this document.

IV. Instrumentation

Some projects will be able to access existing data that will meet the needs of their studies, while others will need to administer an assessment in order to study the effects of the LSC on students. To determine if existing data will meet your needs, you should consider the following questions:

- Are the outcomes that were measured relevant and important in light of the goals of the LSC, the LSC guidelines, and the information needs of your stakeholders?

- Are the instruments valid and reliable?

- Are the instruments potentially sensitive to the LSC treatment?

- Are the outcomes measured in a way that will be acceptable to your key stakeholders?

- Are the data reported at the individual student level, or at least at the classroom level, so you will be able to design a reasonable study?

An obstacle faced by many LSC projects is that state- or district-mandated assessments are often not aligned with the goals of the LSC. One option, as described below, is to administer an additional assessment that is aligned with the project's goals to all the students involved in the project or to a sample of classes. A second option is to construct a sub-scale that is fairly well-aligned with the goals of the LSC. This option is feasible only if you have access to results for individual items on the assessment and access to the assistance of someone knowledgeable about measurement issues.

There is no one way to determine alignment, and a number of different approaches can be used. You may find it helpful to review the work of Norm Webb and see how his approaches might be used in your project. Information on alignment between expectations and assessments can be found in several articles by Webb, which are available at the following websites:

- http://www.wcer.wisc.edu/nise/Publications/Briefs/Vol_1_No_2/

- http://www.wcer.wisc.edu/nise/Publications/Research_Monographs/vol6.pdf

- http://www.wcer.wisc.edu/nise/Publications/Research_Monographs/vol18.pdf

These articles describe the three major methods of alignment—sequential development, expert review, and document analysis. In addition, five categories of criteria for judging alignment are presented—focus of the content, articulation across grades and ages, equity and fairness, pedagogical implications, and system applicability (realistic and manageable in the real world).

If you determine that previously collected data are not useful for studying the effects of the LSC, or if you do not have teacher or student level data, you will likely need to administer an assessment that meets both of these criteria.

There are several issues to consider when selecting an assessment. First, you must consider the information needs of your stakeholders. Second, you need to recognize that the types of assessment tools your project uses will convey some messages about what you think should be taught and learned. In selecting an assessment tool, it is important to balance practical concerns—what is or might be easily available—with what that choice might say about your teaching goals. Traditional measures include, for example, scores on routinely administered achievement tests such as the Iowa Tests of Basic Skills (ITBS), Tests of Achievement and Proficiency (TAP), the ninth version of the Stanford Achievement Tests (Stanford-9), the Metropolitan Achievement Tests (Metro), the Comprehensive Tests of Basic Skills (CTBS), or Terra Nova. Advantages to this approach include its familiarity to most of the stakeholders and the availability of many commercial products. Indeed, the data from such instruments may already be available through your local or state assessment programs. A disadvantage is that these products tend not to model the kind of teaching and learning embraced by reform efforts; although multiple-choice tests can assess reasoning and higher order thinking skills, the tests currently available at the elementary and secondary levels rarely do so.

Another approach, and one in-line with recommendations in the national science and mathematics standards, is to use open-ended items or performance assessments that involve multiple responses that can reflect real-life, complex problems. Developing such measures can be a challenge, however, and many assessments which appear to be valuable because of their "authenticity" might have questionable reliability and validity (e.g., they focus on a very small subset of the domains of interest), limiting the extent to which the results can be generalized. Disadvantages of this approach include the difficulty of finding an appropriate instrument and the amount of time needed to administer and score the performance items, as well as the costs associated with each.

Often, commercial tests, regardless of whether they use multiple-choice items or performance tasks, include multiple sub-scales. For example, the New Standards Reference Exam reports student performance as an overall mathematics score and on the following sub-scales: skills, concepts, and problem solving. By analyzing student scores on the sub-scales you might be able to address the information demands of different sets of stakeholders. The reform community might care most about student performance on the concept and problem-solving scales while other stakeholders might be most interested in knowing how students performed on the skills scale. When choosing an assessment, you should consider which, if any, sub-scales the instrument contains.

You might want to use both multiple-choice and performance assessments to take advantage of the benefits of each. Both the New Standards Reference Exam and the assessments developed by the Partnership for the Assessment of Standards-Based Science (PASS) include both multiple choice items and performance tasks. Similarly, the Iowa Tests of Educational Development (ITED), a commercial product for grades 9–12, requires students to read long passages, solve multi-step mathematics problems, and analyze simulated science experiments. Other instruments, including the Stanford–9 and the Terra Nova, also have optional performance assessment components. Whatever test(s) are selected, you should be sure that they:

- Are valid—the assessments measure what you intend for them to measure;
- Are designed for the population you will be assessing;
- Are reliable—if a student takes the assessment multiple times, his/her score will remain stable; and
- Report scores at a level (teacher or student) that will give you enough cases to conduct a meaningful analysis of the data.

Another issue to consider is the metric on which outcomes are reported. Two types of approaches are used most frequently for conceptualizing student outcomes: continuous scales and categorical scales. Continuous scales provide data on changes occurring over some range of possible outcomes, such as percentile ranks or normal curve equivalents, both of which operate on a 0–100 point scale. Categorical scales are far more restricted, employing only a few gradations. Proficiency scores, reflecting below-, at-, or above-level expectations, provide a categorical outcome metric that is very popular today. Both approaches have their pluses and minuses. Proficiency scores send the message that all students are expected to reach the same high standard, but do not measure growth within a proficiency level. On the other hand, continuous scores highlight improvement and allow you to examine more finely grained changes in student achievement. However, relatively small changes on continuous scales may be statistically significant when large samples are used but educationally meaningless.

Commercially-available tests frequently include both types of outcome metrics (continuous and categorical) in order to meet a variety of user needs. Careful thought needs to be given to the selection of an outcome metric within each individual study, as the outcome metric plays a big role in determining the types of statistical analyses you can apply, as well as the types of conclusions you can draw from your study.

V. Sampling

Most states and districts assess mathematics achievement every year, at least in selected grades, and some assess science as well. If the project decides that one of these instruments, or some definable subset of an instrument, is an adequate measure of the LSC goals, then sampling is not usually an issue. If you can get scores on all students, you probably will want to use them all in order to maximize your ability to detect gains. On the other hand, if your study design calls for linking additional information to student records, e.g., the number of hours of professional development each student's teacher has had, you may need to work with a sample of the student records. Similarly, if you need to get signed releases from parents in order to access the scores, you might want to select a sample of students.

More typically, however, sampling comes into play only when (1) you need to administer an assessment instrument either because test scores are not available, or because the tests being used are not appropriate for your goals; and (2) it is too expensive to administer and score an assessment for the entire student population at one or more grade levels.

A sampling textbook will tell you that a simple random sample is likely your best bet, where you might make a list of all students whose teachers are participating in the project and then select every 2nd, 5th, or 10th name, depending on the total size of the sample you need to have in order to have a reasonable chance of detecting growth due to the LSC. But the realities of school life typically make this strategy infeasible; teachers will not take kindly to your pulling random students out of their classes. A more feasible alternative is to make a list of classes and randomly select from those to get the number of students you need, enabling you to administer the assessments to intact classes.

Generally, the larger your sample, the more likely you are to be able to show gains. For example, a study with 100 students (50 treated and 50 untreated) would have a reasonable (80 percent) chance of detecting a difference between the two groups of half a standard deviation. A sample size of 300 total would virtually ensure that you could detect a difference of this magnitude. In contrast, a study using a sample of only 50 students total is far less likely (a 54 percent chance) to be able to distinguish a difference of this magnitude from random noise. Looking for smaller gains requires even larger samples, e.g., you would need over 1100 students to have an acceptable chance (80 percent) to detect a gain of half this size.

In selecting your sample, you need to decide to which population you want to generalize. If your LSC covers the K-8 range, you need to make sure your sample covers that range in some reasonable way. If this is not possible, you must delimit either the claims you make from your findings or present a convincing argument for drawing broader conclusions.

In quantitative studies, the sample should be large enough that the study has sufficient statistical power to detect meaningful differences over time and/or among comparison groups. In qualitative studies, trustworthiness of the research is an issue, since the sample is likely to be small. In selecting a sample for qualitative studies, it is important to be able to make the argument that the results are not isolated examples arising as a product of selective sampling of data or subjects.

In designing your study, tradeoffs are often necessary. You need to use data collection approaches that are feasible and cost effective. A brief example of the types of choices you will need to make follows:

> Project XYZ has high quality performance tasks but can only afford to have 200 students' work analyzed by expert scorers. They have several options, including:
>
> a. Administer the tasks to all classes, have teachers score their own students, and compare the results of classes where teachers have participated in varying amounts of professional development;
> b. Same as "a," but have teachers score each other's classes;
> c. Administer the tasks to 100 randomly selected students of each treatment level (heavy and low/no treatment); or
> d. Administer the tasks to all classes and randomly select 100 students of each treatment level to score by an expert.

While all of these options are subject to selection bias (which is discussed later in the handbook), if the highly treated teachers are different from the less treated teachers, some alternatives are better than others. Choices "a" and "b" likely have the problem of unreliability of scores unless the teachers have had substantial training in scoring. Choice "a" also has a problem regarding the apparent lack of objectivity. Choice "c" would create the feasibility problems of pulling individual students out of classes. Given the constraints, choice "d" is probably the best option; it is feasible in terms of both cost and administering the test to whole classrooms, sends the right message to teachers about the objectivity of the scorer, and can be used as a vehicle for professional development.

VI. Design

The design of your study is the foundation for a number of decisions about what data you need to collect, the analysis that you will do on that data, and the conclusions that you will be able to draw as a result of your study. For LSC student outcome studies, you want to be able to determine if student outcomes have changed as a result of the teacher enhancement project. Your design, therefore, should allow you to show both that student outcomes have changed, and that any changes in student outcomes are likely a result of the LSC project and are not primarily due to some other factor. A strong study design can increase the chances that you measure a true effect, that is, your project caused the change. The integrity and credibility of your conclusions depend on having an appropriate and sound study design.

There are two fundamental design features that your study should include in order to make the case for your LSC's effect on student outcomes. First, the study is best if it involves a comparison or control group in addition to a treatment group. In cases where a comparison group is unavailable, a comparison to an accepted standard (e.g., outcomes of students in similar districts or grade-level equivalent scores) might suffice. Second, the study should examine the initial status of the comparison or control and treatment groups, so that you can make a case that they were initially equivalent or adjust for initial differences.

A. Using a Control or Comparison Group

Consistent with good research design, studies of the effect of the LSC on student outcomes will be strongest if they include both a treatment group and a comparison group. You can accomplish this by comparing outcomes for students of teachers participating in the LSC to outcomes for students of teachers not participating in the LSC, or by comparing outcomes for students of teachers fully participating in LSC to outcomes of students of teachers with more limited involvement. Using a control or comparison group allows you to examine the effect of the project's treatment aside from other factors. Without a control group, it is nearly impossible to say that any change in outcomes is due to the treatment as opposed to other factors.

As an example, consider a school district that is shopping for a new reform-oriented elementary school mathematics program. They have narrowed their choices down to two programs. To help make the final decision, the school board asks the publishers to present evidence that their mathematics program helps students learn mathematics.

The first publisher shows the school board results of a study that compared 3rd graders' test scores before and after using that curriculum package. The results of the study show that the 3rd graders significantly improved their mathematics scores on the SAT–9 over the course of 1 year.

The second publisher then presents the results of their study. They also found that students scored significantly higher in mathematics on the SAT–9 after 1 year of exposure to their curriculum (comparing end of grade scores at 2nd and 3rd grades). Further, this increase was significantly larger than the increase in scores of students in the same school district exposed to the traditional mathematics program, which also is used currently by this district.

Which program do you think the school board should adopt? While the first publisher's study showed that students achieved more after using their program, there is no evidence that the increase in student scores is attributable to their mathematics program. After all, the students are older and have been in school for an entire year between the two administrations of the assessment. Most likely, the students would have scored better on the assessment after one year if they had experienced *any* mathematics instruction, perhaps even if they had experienced *no* mathematics instruction. Without further supporting data, this type of design can be readily challenged.

In contrast, the second publisher's study used a control group to eliminate those possible explanations for the change in mathematics scores. Although the second publisher's study does not answer every question one might have about the curriculum's effectiveness, it clearly makes a stronger case that their curriculum is more successful than the traditional one at helping students learn mathematics.

It is sometimes difficult to construct comparison groups with no exposure to the treatment, especially if a project is directed at a whole school or district. It is frequently more feasible to use data on documented differences in level of treatment as a way of defining groups for comparison in a study. If all teachers in your district are participating to some extent in the LSC program, you might look at the amount of training each has received as a way of defining your treatment and comparison groups. For example, you could examine the scores of students whose teachers had participated in the program for three years compared to students whose teachers had participated for only one year. Please note, however, that the appropriateness of this approach would depend on how teachers are selected for training. If teachers could volunteer to participate in the first year and only the stronger teachers tended to do so, it would not be appropriate to use teachers' years of treatment as a way to select groups for comparison, because the teachers who volunteered early in the project might well have been different from those who did not, even prior to the LSC.

Another approach would be to create groups for comparison based on the degree to which teachers are implementing the instruction intended by the project. For example, you could determine the number of instructional units used by each teacher or use a measure of the extent of implementation of intended instruction based on questionnaires or classroom observations. This approach, however, again runs the risk that teachers who are able to implement the project well were different in important ways prior to the LSC from those who are not able to do so.

Sometimes no comparison group is available, e.g., if all teachers participated in the same professional development activities at the same time. In these cases, conducting pre- and post-tests with the treatment group might still be considered if you can compare any change to an expected level of growth or to changes in a similar population. However, this approach must be backed up with a very solid logical argument that changes were the result of the treatment. This logical argument could be strengthened through the conduct of a supplementary mini-study. For example, let's say that the LSC professional development training consisted primarily of a one-week intensive program held during the summer. Although all teachers were required to attend, five were not able to do so due to illness or family matters. It might be possible to have the students of these teachers serve as a control group that can be compared to a similar subset of the treatment group.

B. Examining Initial Status

The second key element to a strong research study is examining initial status of the treatment group and the control or comparison group(s), in order to establish equivalence of the groups prior to the treatment or to take initial differences into account when drawing conclusions. This requirement can be accomplished in a variety of ways:

- By using repeated measures of the outcome, generally a pre-test and post-test;

- By using a relevant covariate to adjust for initial differences, if necessary;

- By using matched samples; and

- By making a reasonable case that the groups are initially equivalent.

The most straightforward way to examine initial differences among groups in a study is to use data on the outcome variable prior to treatment. By including such data in a study, you can examine changes in the outcome variable directly and determine if there are differences across groups. Generally termed a pre-test, post-test design, this approach requires that the outcome be measured more than once for each group.

If data on the outcome of the study that measure initial status prior to treatment are not available, other data closely related to the outcome can be used as an alternative. For instance, student reading achievement scores tend to be highly correlated to mathematics achievement scores. These kinds of related data are termed covariates. Ideally, measurement of the covariates would occur prior to treatment, but measurement of a covariate during or following treatment is acceptable as long as the covariate is not likely to be affected by the treatment. A covariate may be used either to show that groups were initially equivalent or to adjust for initial differences.

A third approach to examining initial equivalence or difference among groups is to use matched samples. Doing so requires information about characteristics of the groups that might be related to the outcomes being studied, and, if not controlled, might offer explanations other than treatment as part of the LSC for differences in outcomes. For example, characteristics such as race, socioeconomic status (free or reduced-price lunch eligibility is frequently used), gender, and ability should be fairly consistent across the groups being compared. While matched sample designs do not typically include initial equivalence or difference information in the outcome analysis, they do minimize the likelihood that initial differences were present. The more alike the groups were initially, the more likely it is that any measured difference in outcomes is due to the one characteristic that is known to be different, namely treatment in the LSC. The major difficulty with matching is that you can never be sure that all the relevant factors were considered that might be critical in explaining differences across the groups.

A fourth approach, clearly the weakest of the four, is when, in the absence of data to show the extent of similarity, you try to make a reasonable case that the groups were initially equivalent. The logic of this approach is similar to that of the matched samples approach, but it differs in that data about important group characteristics are not available. For example, the treatment and comparison groups might have come from schools in similar districts or might have come from the same schools, but from different teachers. It clearly would be better to know more about the characteristics of the groups, but at least some argument is provided that the group being treated is not otherwise dissimilar in important ways from the comparison group.

Study Designs: Four Examples

Different research designs incorporate none, some, or all of the necessary elements for a defensible study of the LSC's effect on student outcomes. Four of the most common designs in education research are discussed. For each design, an example is presented and the strengths and weaknesses are identified. Although the last design is clearly the strongest, it is possible to enhance the other designs, in effect making them more like the last design.

A. Treatment Group Only, Post-Test Only

In this design, one group of students is observed or tested only after they have received the treatment (instruction from an LSC treated teacher). No information is available on the level of treatment of these teachers in the LSC program. For example, an elementary science LSC might have student scores on a district assessment in science given in the fourth grade. However, these data are not linked to the students' teachers, nor to any prior science achievement score or other potential covariate.

The scores on this science assessment show how students in the district are performing in science at the fourth grade level. This information is useful for examining student mastery of certain skills or concepts (similar to what teachers seek in their classrooms when they administer end-of-unit assessments). However, the design gives you little or no chance to demonstrate a relationship between the LSC and the student outcomes, because you cannot show growth on the outcome over a period of time in which the LSC might have influenced scores. Further, you can not judge whether the LSC treatment had any effect on the outcome; the students may have scored better, worse, or the same with or without the LSC.

There are several reasons why this design does not allow you to make the case that the LSC had an effect on student outcomes. The design lacks any comparison groups, either treated or untreated groups or groups that differ in their levels of treatment, nor is there any comparison to a standard of achievement. Since there are no comparison groups, the design cannot examine initial status differences among the groups. In fact, the initial status of the students prior to the LSC is unknown; the scores may be the result of little or no change in science achievement or a very large change. For these reasons, this design fails to provide evidence of the effect of the LSC on students, and its use is discouraged. It is presented primarily to show how the other designs address some of the deficiencies with this design.

B. Treatment Group Only, Pre- and Post-Tests

In this design, students are given a pre-test or baseline measure, then the treatment (instruction by a teacher targeted by the LSC), and finally a post-test. The pre-test and post-test scores can be compared to examine growth. Note that this design, in its basic form, does not include any comparison groups. An example of this design follows:

> A group of secondary mathematics LSC teachers all used a previous end-of-course exam in Algebra as a pre-test for their Algebra students at the beginning of

the school year. At the end of the year, they

> compare their students' results on the pretest to their scores on this year's district-mandated end-of-course exam in Algebra. The teachers can link individual student scores between the two tests by the students' names, so the change in test score from the beginning of the year to the end of the year are available. The comparison shows significant growth in Algebra achievement for these students.

This design contains some of the elements of a strong study, but falls short on others. The design includes a measurement of initial status on the outcome of interest for only one group, students whose teachers received LSC training, but does not include the use of a comparison group. The growth in Algebra achievement is known for the treatment group, but how that growth would compare to a similar group of students receiving a year of Algebra instruction by non-LSC-treated teachers is unknown. The effect, therefore, is difficult to attribute to the LSC professional development. Further, if the same test is used for pre- and post-testing, it is possible to argue that the test itself caused the change by sensitizing students to what was important to learn.

The study would be strengthened by providing a comparison group. It should be noted, however, that even if an appropriate comparison group of students is used, the teachers of those students still might not be comparable to the LSC–trained teachers. It could be the case that the LSC teachers have been the ones whose students always performed very well on the end-of-grade Algebra exam, even prior to the LSC. Sometimes, in such cases, statistical adjustments can be made to make these groups equivalent. However, if the proposed comparison group is considerably different from the control group, it probably is not an appropriate group to use.

C. Treatment and Control Groups (or Varying Levels of Dosage), Post–Test Only

In this design, you either have two groups of students, with only one group receiving instruction from teachers participating in the LSC, or groups of students who receive instruction from teachers with varying amounts of participation in the program. In this design, all groups are tested once, after a period of treatment, and their outcome scores are compared. The following example illustrates this two-group, post-test only design:

> A K-8 science LSC has been taking place in several districts, one of which is testing science in the 8th grade in a standardized way for the first time this year. The LSC perceives this as an opportunity to

> study the contribution of the LSC to the students' scores on this assessment. The LSC has information on the extent to which teachers have participated in LSC activities over the three years of the project. District records allow the LSC to identify which science teacher each student in the district has had in the past three years. Students' scores on the science assessment, therefore, can be grouped by the number of years the students received instruction from a teacher who had participated in the LSC for more than 30 hours.
>
> The results show that with each additional year of instruction from an LSC-trained teacher, students performed better on the science assessment.

Like the previous design, this one contains some features of high-quality research. The design of this study includes comparison groups so that differences in outcomes across groups can be examined directly. However, examination of initial status, particularly possible initial differences across groups, is not included. In order to strengthen this study, at least one of the methods for examining initial status should be employed.

D. Treatment and Control Groups (or Varying Levels of Dosage), Pre- and Post-Tests

In this design, you have two groups of students, a treatment group and a control group. A pre-test and a post-test are administered to both groups of students. Overall, this is the strongest design presented as it includes both a comparison group and the means for examining initial equivalence of the two groups. Further, if the groups are not initially equivalent, the pre-test scores allow you to make an adjustment in your analysis for the initial difference. Consider the following example:

> A mathematics LSC is located in a state that mandates end-of-year tests in reading, writing, and mathematics for all students in grades 3–8. The district also uses electronic cumulative folders allowing them to track students' progress over time. Using this information, the LSC analyzes students' growth, as measured by the change in test scores year-to-year, by the number of years the students had an LSC trained teacher. The district thus is able to create three groups of students, those with 0, 1, and 2 years of instruction by a LSC-trained teacher. The analysis shows that students with one year of

> instruction by an LSC-trained teacher had larger gains in their mathematics scores than students who never received instruction from an LSC-trained teacher. Students with two years of instruction from an LSC-trained teacher had larger gains than each of the other two groups.

This design is the strongest of the four presented and allows for a credible case to be made that the LSC is responsible for the differential gains of the students. It is important to note that, while the use of a pre-test allows for any initial differences among students to be controlled through use of an analysis of covariance, differences among teachers are not controlled. Thus, as described in the next section, you need to examine your groups for possible selection biases.

VII. Strengthening Your Study's Internal Validity

Studies that are assessing the effect of an educational program must give strong consideration to internal validity. Internal validity means that you have evidence that your program, and not other factors, was the cause of the outcomes. Such alternative explanations are known as threats to internal validity. A good study design helps to minimize these factors, but even the best studies have potential threats to internal validity. Thus it is the responsibility of the researcher to examine the study for any threats and to determine the likelihood that the threat, and not the treatment, was responsible for any differences in the outcomes.

In studies using comparison groups, the largest potential threat to internal validity is related to sample selection—that the treatment and control groups are selected in different ways, resulting in bias. For example, in districts with high teacher turnover, untreated teachers might tend to be new to the profession while treated teachers would tend to be more experienced. Thus, there would be an inherent bias in a comparison of these two groups of teachers. For this reason, it is critical that you build into your research design some method to examine the initial equivalence of your treatment and control groups.

The following example, while exaggerated, illustrates the threat of selection bias. Imagine we wanted to investigate the effect of taking calculus on mathematics achievement, with the hypothesis that students who take calculus will be better prepared in mathematics than students not taking calculus. To do this, we examine students' scores on the mathematics portion of the SAT relative to their score on the PSAT, comparing those who took calculus to those who did not. The results of the analysis show that those students taking calculus have much greater gains than the students not taking calculus. While taking calculus may lead to higher gains between the PSAT and the SAT, this study does not justify enrolling everyone in calculus in an attempt to raise mathematics achievement. Rather, it is very likely that the calculus students would have higher gains than the non-calculus students even if they hadn't taken the course, since the students who elect to take calculus tend to have a particularly high capacity to learn mathematics.

In their seminal work on research design, Campbell and Stanley identify eight threats to internal validity that could interact with the selection of your treatment and comparison groups. Of those, the four threats you are most likely to encounter in research on the effects of the LSC are:

- History – students in one of your groups have an experience other than what your teacher enhancement program provides. For example, an exciting new, museum-based science education program for elementary children advertised in LSC professional development sessions might be attended primarily by students of treated teachers. That program rather than the LSC treatment could be the primary reason for the difference in test scores between the groups.

- Maturation – a change occurred simply as a result of the passage of time. For example, if the students in your treatment group are more advanced before the LSC treatment is implemented, they might develop at a faster rate than students who start off at a lower level.

- Statistical regression – where those who score very high and very low initially have a tendency to score closer to the mean. Therefore, you want to make sure that students at the extreme ends of your measurement scale are not concentrated in one of your groups.

- Experimental mortality – there is considerable attrition in the study, particularly if participants in the treatment and control group drop out at different rates.

Without random assignment of teachers to treatment groups and students to teachers, neither of which is typically feasible, there is no research design that can totally rule out these threats. However, there are three common methods researchers use to help reduce the possibility that these threats to validity are responsible for the study's outcome:

- By argument – This is the easiest action but is also the weakest. If you have knowledge about the students and teachers in the groups, or about how they were selected, you can make the case that there was or was not a selection bias.

- By measurement or observation – Sometimes you can measure the threat in order to subtract it out. For example, if you find that most of the students in your treatment group also are participating in an after school program, you could compare their test scores to other students in the after school program—not in LSC classrooms—to measure the effect of the after school program.

- By analysis – Some threats can be addressed by advanced statistical analysis. Examples include computation to adjust for regression effect and analysis of variance (ANOVA) to adjust for mortality.

Regardless of how you choose to examine the possibility of selection bias in your study, you need to address this issue fully in your report.

VIII. Analysis

The credibility of your study will be bolstered through a sound analysis. The analysis tools that you use in your study need to be consistent with your study design, the type of outcome data that you are including, and the levels of data (student, teacher, etc.) that are represented in your study. Ultimately, the analysis should allow you to determine whether the student outcomes under study have changed, and whether any change is likely to be a result of the LSC. It is important that you make a case for the appropriateness of your analysis based on these concerns.

For a quantitative study, regardless of whether your outcomes are measured on continuous or categorical metrics, your analysis will include two main phases: describing your data and statistical testing. Issues involved in each of these, as well as disaggregating data by demographic subgroups and analysis of qualitative data, are discussed in the following sections.

A. Descriptive Statistics

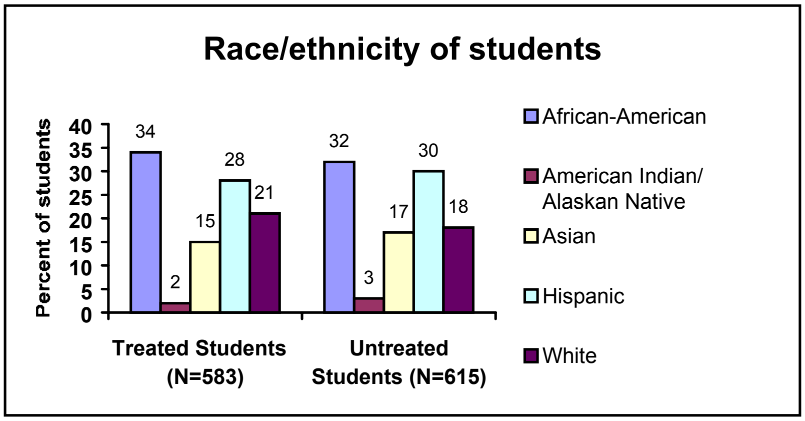
Descriptive statistics are used to provide simple summaries of your data. They form the basis of most quantitative analyses of data and are frequently illustrated with simple graphic displays. In general, you want to report descriptive information about the students in your study (race/ethnicity, gender, etc.) both overall and for each treatment group. You also might want to disaggregate your outcome measures by race/ethnicity, gender, or some other demographic characteristic of interest to look for differential performance among certain sub-populations (see section C below).

The appropriate method for describing your data depends solely on whether the data are measured on a categorical scale (e.g., demographics, attainment of standard) or continuous scale (e.g., percentile score). If the data are categorical, it is typical to report the overall number of students and the percent of students in each category. For example, you may choose to present the race/ethnicity of students in your treatment and comparison groups as follows:

Percent of students

Race/ethnicity	Treated N = 583	Untreated N = 615
African-American	34	32
American Indian/Alaskan Native	2	3
Asian	15	17
Hispanic	28	30
White	21	18

Another option would be to present the data using a bar chart or histogram:



When presenting continuous data, it is appropriate to report measures of central tendency and dispersion as well as the number of cases in each group. In most cases, you will want to use the mean as your measure of central tendency, rather than the median or the mode.

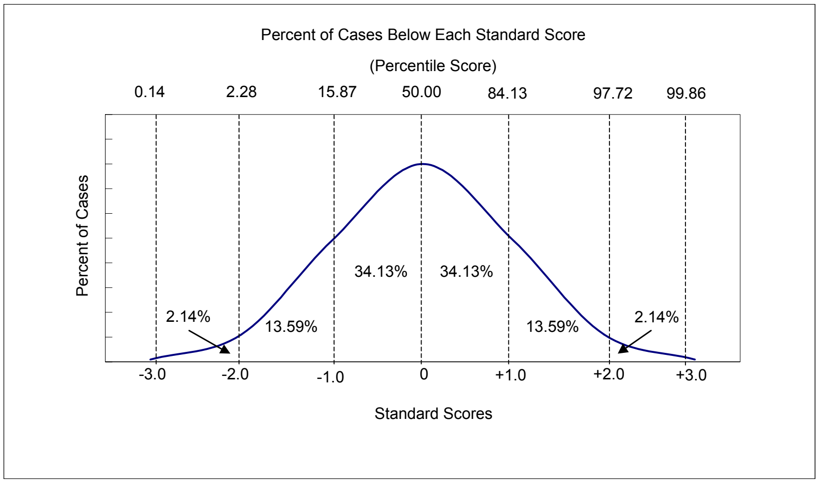
While the mean provides a considerable amount of information about your data, it is generally not a sufficient descriptor. You also need to indicate how the data are dispersed around the mean. The most useful estimate of dispersion is the standard deviation. The formula for the standard deviation is based on the distance that each score is from the mean, and it is usually calculated using statistical software such as SPSS or SAS. If your scores have a normal distribution—a bell-shaped curve or something close to it—the following statements can be made:

- About 69 percent of the scores fall within one standard deviation of the mean (this includes the area both above and below the mean);

- About 95 percent of the scores fall within two standard deviations of the mean; and

- About 99 percent of the scores fall within three standard deviations of the mean.

Classic Bell-shaped Curve

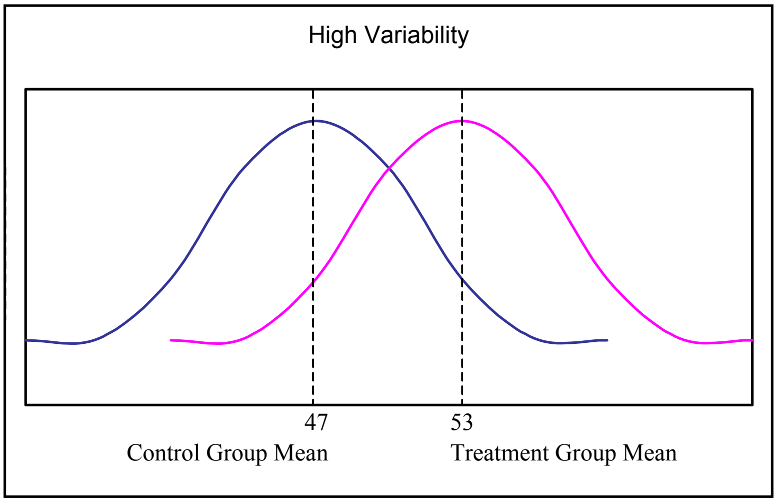


B. Inferential Statistics

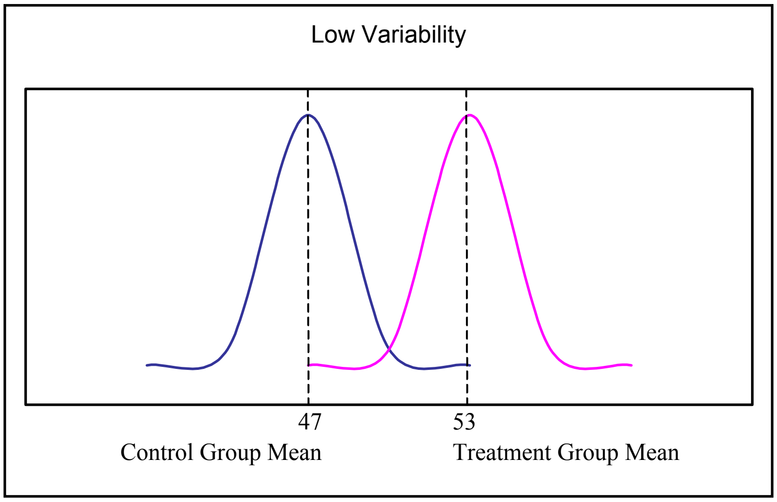
Inferential statistics are used to test whether the data you obtained from your sample reflect the results that would be obtained if you used the entire population. Inferential statistics also are used when you want to show that the difference between your treatment and control groups is dependable and not the result of chance. Common inferential statistical procedures are the t-test, chi-square test, analysis of variance (ANOVA), regression, and hierarchical linear modeling (HLM).

For example, you may want to compare test scores of your treatment group (students whose teachers had participated in the LSC project) with your control group (students whose teachers did not participate). As a part of your analysis, you have already calculated the means for the two groups. Let's say that the mean for the treatment group is 53, and the mean for the control group is 47. Since the means are different by 6 points, it might appear that the two groups are different, but it is not quite so simple. The means alone do not give enough information; you also need to know about the dispersion around the mean. This concept can be illustrated by using two extreme examples. Notice that in both examples the mean for the control group is 47, and the mean for the treatment group is 53. However, in Example A, the scores of each of the two groups (treatment and control) vary a great deal, and the scores for the two groups overlap a great deal. Example B provides quite a contrasting picture in which the scores within the treatment and control groups show little variation, and there is little overlap between the two groups. The two groups appear to be most distinct in Example B.

Example A:



Example B:



There are several factors that influence the selection of the appropriate statistical test. Most important are the metrics upon which the outcomes are measured (continuous or categorical) and the design of your research study. However, this is not all you must consider in this decision. All statistical tests are based upon assumptions about the data. For example, the t-test assumes that your data are normally distributed, and the ANOVA requires homogeneity of error variances. If a statistical test's assumptions are not met, the test could give spurious results and another statistical procedure should be considered (e.g., rank-ordered comparisons). For this reason, it is often advisable that you seek the help of a knowledgeable statistical consultant for the analysis of your data.

Because statistical tests are sensitive to the number of cases in your groups, tests involving a large number of cases often detect as statistically significant differences that are too small to be of practical significance. Thus, it is desirable to include a measure of magnitude of the difference between your groups—commonly referred to as the effect size. While the statistical test can tell you if the difference between your groups is significant (likely real rather than due to chance), the effect size helps you decide if the difference is substantial. When comparing means, the effect size is the number of standard deviations between the means of the two groups (i.e., the difference between the means divided by the overall standard deviation). Typically, an effect size of 0.2 is considered small, 0.5 medium, and 0.8 large.

C. Disaggregating Data by Demographic Subgroups

You might find that the treatment group as a whole does significantly better on the post-test than the control group. However, some students may be experiencing a greater effect than others. Conducting separate analyses for various demographic subgroups will show if there are patterns in the outcomes; this process is called disaggregating the data. For example, you might find that boys are scoring considerably higher than girls and subsequently would want to explore the reason for the differences, so you could take appropriate steps to ensure that all students are having an opportunity to learn important science/mathematics. Common subgroups for disaggregating the data are gender and race/ethnicity. Other subgroups that you might consider are class size or number of students in poverty as determined by their participation in the free and reduced-priced lunch program (sometimes easier to do at the elementary than secondary levels). One issue in doing this type of analysis is that you must have enough students in each group. Indeed, a rule of thumb is that you should have a minimum of 20 cases in each subgroup for analytic purposes.

D. Qualitative Analyses

Frequently it is useful to combine quantitative analysis with qualitative analysis. The former provides an overview of success as determined by outcomes that lend themselves to direct measurement and numerical summarization; the latter provides information on outcomes that are best addressed through rich description.

For qualitative analyses, it is important to provide full descriptions of how the data were collected, how the data were analyzed, and how conclusions were drawn from the analysis. Depending on the number of cases you include, either individual case studies or an integrative analysis across data sources might be reported. If the latter approach is chosen, examples from the data should be provided to support your methods and your conclusions. Such examples will both enable your audiences to judge the credibility of your conclusions and gain a deeper understanding of the context of the effects observed.

IX. Reporting

The strength of your study depends heavily on appropriate decisions related to all of the components of the study that have been addressed so far. Yet, even the strongest study will be of very little interest or value if it is not reported appropriately and well. At the same time, a variety of stakeholders will be interested in the results of your project, and how you report your results will often vary depending on the audience. NSF and the research community will be interested in the technical details, while a summary of the project will generally be more appropriate for the school board and parents. Thus, you should be prepared to develop more than one report, each appropriately presented for its audience.

A full report should provide an overview of the project, including the goals and objectives. You should indicate why you are studying the student outcomes you have chosen, and how they are connected to the goals of the LSC and your teacher enhancement project.

It is extremely important that you provide a clear description of your study. The description should build a case for the approach you took by describing the instruments used, sample characteristics and selection, study design, and choice of analyses. Give sufficient detail so that the reader can understand and judge the credibility of the analyses undertaken. It is not enough to say, for example, that the treatment and comparison groups consisted of students in the 4th grade. Data about these students, such as their prior achievement levels, socioeconomic status, gender, and race also should be provided. Information about any special programs, in addition to the LSC, in which they, their teachers, or their schools might be participating, also should be provided.

The report should indicate the number of students involved and a rationale for any sampling decisions. Groups included in the design should be specified. The report also should include, for all groups and all variables, either frequency distributions for categorical data or means and standard deviations for continuous data. Inferential statistical analyses, including the test statistic, degrees of freedom, p-value, and effect size should be presented as well. Tables and graphs often help to organize and explain the data succinctly; they also help to communicate the most important results in meaningful ways.

Alternative explanations for any differences detected in the study results should be identified. If further analyses ruled out these explanations, these analyses should be presented, and if alternative explanations remain, they should be acknowledged and reasonable arguments regarding their likelihood presented. Conclusions, implications, and generalizations that can be drawn from your study also should be provided. It is also important to offer any lessons that you have learned that might be applicable to related efforts so that others can benefit from your experience in conducting the study, as well as from your results.

Reports for other audiences, such as school boards, principals, or parents, probably would focus on the purpose of the overall project and the specific issues you addressed in your study. These audiences would be extremely interested in the results and implications, which would need to be described in very straightforward terms that speak to the issues of greatest interest to each audience. Most of the details of the instrumentation, sampling, and analysis would not be appropriate for these audiences, although you should be prepared to make that information available upon request.

APPENDIX A

Quick Reference:
Handbook for Conducting Studies of the Effect of the LSC on Student Outcomes

Instrumentation

It is up to each project to decide which instruments to use in measuring student achievement and other student outcomes. Optimally, you would examine student outcomes using multiple measures in order to satisfy the information needs of a variety of your stakeholders, to triangulate findings, and to provide a rich array of evidence of the effect of the LSC on students. Each project should make a case that:

- The studied outcomes are relevant and important to the LSC project;
- The chosen instruments appropriately measure the studied outcomes;
- The instruments are reliable; and
- The instruments are potentially sensitive to the LSC treatment.

Sampling

Studies might include data for all students targeted by the LSC, but more likely will include data from a sample from one grade level and/or from a subset of districts, schools, or classrooms. The studied sample should be:

- Representative of the population of students being targeted by the LSC;
- Exposed sufficiently to the LSC in order to merit an examination of effect; and
- Large enough to provide statistical power to detect differences in outcomes or to ensure trustworthiness of qualitative methods.

Design

The study design should enable the researcher and the audience to answer the question: To what extent has the LSC had an effect on student outcomes? Strong studies:

- Compare outcomes for treated students to outcomes for untreated students; and/or
- Compare outcomes for students with varying degrees of treatment; and/or,
- Compare outcomes of treated students to another standard (e.g., outcomes of students in similar districts or grade-level equivalent scores).

An examination of the initial equivalency of comparison groups on outcomes is usually necessary. Options for examining equivalency, in decreasing order of preference are:

- Using pre and post measures;
- Using a relevant covariate (e.g., reading test scores);
- Using matched samples; or
- Making the case that samples are initially equivalent, or that an appropriate standard of comparison has been chosen.

Quick Reference:
Handbook for Conducting Studies of the Effect of the LSC on Student Outcomes
(continued)

Internal Validity

The credibility of a study can be undermined if alternative explanations for the results, such as selection biases, are ignored. A sound study will:

- Identify plausible alternative explanations for its results;
- Address plausible alternative explanations, either through statistical methods or through arguments with evidence refuting alternative explanations; and
- Acknowledge remaining shortcomings of the study, possibly providing recommendations for further research to address those limitations.

Analysis

Analysis methods and tools should be consistent with the study design and the type and level of outcome data being investigated. An appropriate qualitative analysis describes how the data were collected and analyzed, and how conclusions were drawn. An appropriate analysis in a quantitative study includes both descriptive and inferential statistics.

The goals of reporting the results of your study are to communicate your findings to audiences with an interest in your LSC and to make the case that the results of the study represent the effect of the LSC on the student outcomes you have studied. Including the information outlined here will enable your audiences to judge the study's results fairly.

For instrumentation, a technical report should include the following pieces of information:

- What outcomes are being measured by what instruments;
- Why you expect the outcomes and instruments to be sensitive to the LSC treatment;
- On what metric the outcomes are measured;
- At what level of aggregation the outcomes are reported (e.g., student, classroom); and
- Information about the reliability and validity of the instruments.

For sampling, a technical report generally should include information on the how representative the sample is, including:

- How the sample was selected or determined;
- The size of the sample and of any sub-groups that will be considered in the analyses; and
- Descriptive characteristics of the population, the sample, and any sub-groups.

For design, a technical report should include:

- A description of the design and
- A rationale for the design, making a case that it allows the effect of the LSC on student outcomes to be studied better than reasonable alternative designs.

Analysis (continued)

For internal validity, a technical report should include:

- An examination of possible selection biases;
- Identification of plausible alternative explanations;
- Analysis or evidence to rule out plausible alternative explanations; and
- Acknowledgement of remaining shortcomings.

For quantitative analysis, common conventions for information to be conveyed include:

- Descriptive statistics:
  - For continuous variables, Ns, means, and standard deviations
  - For categorical variables, such as gender and race/ethnicity, Ns and frequency distributions
- Inferential statistics:
  - Test statistic, degrees of freedom, p-value, and effect size; and
  - Information confirming that the data meet the statistical assumptions of the procedures (e.g., normality or homogeneity of error variances).

For qualitative analysis, common conventions for information to be conveyed include:

- How the data were collected;
- How the data were analyzed;
- How conclusions were drawn; and
- Examples from the data.

Reporting

If several reports are produced for different audiences, reference to the most comprehensive report and contact information are generally included so that interested parties can find the most complete information available.

APPENDIX B

Glossary

Analysis of variance (ANOVA)

A test of the statistical significance of the differences among the mean scores of three or more groups. It is an extension of the *t* test, which can handle only two groups, to a larger number of groups.

Baseline

The first phase of research in which outcomes are measured before any treatment is administered.

Bias

Any systematic error that influences the results and undermines the quality of the research.

Categorical scale

A scale that distinguishes among individuals by putting them into a limited number of groups or categories.

Chi-square test

A statistical procedure used with categorical data to test relationships between frequencies in categories of independent variables.

Comparison group

A group that provides a basis for contrast with an experimental group (i.e., the group of people participating in the program or project being evaluated). The comparison group is not subjected to the treatment, thus creating a means for comparison with the experimental group that does receive the treatment. Comparison groups should be as similar as possible to the treatment group but can be used even when close matching is not possible.

Continuous scale

A scale containing a large, perhaps infinite, number of intervals. Units on a continuous scale do not have a minimum size but rather can be broken down into smaller and smaller parts. For example, grade point average (GPA) is measured on a continuous scale, a student can have a GPA of 3, 3.5, 3.51, etc. (See categorical scale.)

Control group

A group that does not receive the treatment. The function of the control group is to determine the extent to which the same effect occurs without the treatment. The control group must be closely matched to the experimental group. (See comparison group.)

Correlation

A statistical measure of the degree of relationship between variables.

Covariate

A variable that a researcher “controls for” in a study by statistically subtracting the effects of the variable.

Degrees of freedom

The number of values that are free to vary when computing a statistic; the number of pieces of information that can vary independently of one another. The degrees of freedom (*df*) tell you the amount of data used to calculate a particular statistic and is usually one less than the number of cases. It is needed to interpret a chi-square statistic or a *t* value.

Descriptive statistics

Statistical procedures that involve summarizing, tabulating, organizing, and graphing data for the purpose of describing objects or individuals that have been measured or observed.

Design

The process of stipulating the investigatory procedures to be followed in doing a certain evaluation.

Disaggregate

To separate data for the purposes of analyses. For example, achievement test scores might be disaggregated to look for separate trends by gender and race/ethnicity.

Dispersion

The amount of variation in the scores around the central tendency. There are two common measures of dispersion, the range and the standard deviation.

Effect size

A statistic indicating the difference in outcome for the average participant who received a treatment from the average study participant who did not (or who received a different level of the treatment). The effect size indicates if the difference is substantial or meaningful. Typically, an effect size of 0.2 is considered small, 0.5 medium, and 0.8 large.

Hierarchical linear modeling (HLM)

A statistical procedure used when data are nested within levels, e.g., students grouped within classes, classes grouped within schools. The method's advantage is that it makes it possible to separate the variance into components explaining the effects of different levels of analysis upon the outcome variable, such as the effects of teacher or school factors on mathematics achievement.

Homogeneity of error variances

An assumption of some statistical procedures (e.g. ANOVA) that the populations from which the samples have been drawn have equal amounts of unexplained variability.

Inferential statistics

Procedures that indicate the probability associated with inferring the characteristics of the population based on data from samples.

Instrument

An assessment device (test, questionnaire, protocol, etc.) adopted, adapted, or constructed for the purpose of the evaluation.

Internal validity

The extent to which the results of a study can be attributed to the treatment rather than to flaws in the research design. Internal validity depends on the extent to which extraneous variables have been controlled by the researcher.

Matched samples

An experimental procedure in which the subjects are divided, by means other than random assignment, to produce groups that are considered to be of equal merit or ability. (Often, matched groups are created by ensuring that they are the same or nearly so on such variables as sex, age, grade point averages, and past test scores.)

Measures of central tendency

The central tendency of a distribution is an estimate of the "center" of a distribution of values. There are three major types of estimates of central tendency: mean, median, and mode.

Normal distribution

An ideal distribution that results in the familiar bell-shaped curve, which is perfectly symmetrical. A large part of inferential statistics rests on the assumption that the population from which we are sampling is normally distributed. The results of a number of statistical procedures are invalid if this assumption is grossly violated.

Performance assessment

A method of evaluating what skills students or other project participants have acquired by examining how they accomplish complex tasks or the products they have created (e.g., poetry, artwork).

Population

The total group of individuals from which a sample is drawn.

p value

Probability value. Usually found in expressions such as $p < .05$. This means "the probability (p) that this result could have been produced by chance is less than (<) 5 percent (.05)." The smaller the number the more likely the result was not due to chance.

Qualitative analysis

The approach to evaluation that is primarily descriptive and interpretive.

Random sampling

Selecting subjects from a population so that every individual subject has a specified probability of being chosen.

Rank-ordered comparisons

Analyses that show the degree of relationship between two variables that are measured on an ordinal scale, that is, items on the scale can be put in order, or ranked, but the intervals between the ranks may not be equal.

Regression

A set of statistical techniques that allow one to assess the relationship between independent and dependent variables.

Regression effect

The tendency for extreme scores to become closer to the mean score on a second testing. Also called "regression to the mean."

Reliability

Statistical reliability is the consistency of the readings from a scientific instrument or human judge.

Repeated measures

A research design in which participants are measured two or more times.

Sample

A subset of a population.

Selection bias

Any factor other than the program that leads to post-test differences between groups.

Stakeholders

Persons who have a vested interest in a project.

Standard deviation

A measure of the spread of a variable based on the average amount that the scores in the distribution are different from the mean. The more widely the scores are spread out, the larger the standard deviation.

t test

A test of statistical significance, frequently of the difference between two group means.

Threats to validity

Factors that can lead to false conclusions.

Treatment group

The group that receives whatever is being applied by the project that distinguishes it from the comparison group.

Triangulation

In an evaluation, it is an attempt to get a fix on a phenomenon or measurement by approaching it via several independent routes. This effort provides cross-validation of results.

Validity

The extent to which an instrument measures what it is intended to measure.

Sources

National Science Foundation. (1993). *User-Friendly Handbook for Project Evaluation: Science, Mathematics, Engineering and Technology Education.* NSF 93-152. Washington, DC: NSF.

National Science Foundation. (1997). *User-Friendly Handbook for Mixed Method Evaluations.* NSF 97-153. Washington, DC: NSF.

Scriven, Michael. (1991). *Evaluation Thesaurus* (4th ed.). Newbury Park, CA: Sage.

Schumacher, Sally, and James H. McMillan. (1993). *Research in Education: A Conceptual Introduction.* New York, NY: Harper Collins College Publishers.

Trochim, William M. *The Research Methods Knowledge Base.* **http://trochim.human.cornell.edu/kb/index.htm**

Vogt, W. Paul. (1999). *Dictionary of Statistics & Methodology: A Nontechnical Guide for the Social Sciences* (2nd ed.). Thousand Oaks, CA: Sage.

APPENDIX C

Additional Resources

Designing research to study the effect of the LSC on student outcomes is not a trivial task. This handbook raises many of the key issues, but it is not possible, nor was it intended, for this document to treat all of the issues at the depth required to transform a neophyte into an expert. Listed below are additional resources on research design.

Bond, Sally L., Sally E. Boyd, and Kathleen A. Rapp. (1997). *Taking Stock: A Practical Guide to Evaluating Your Own Programs.* Chapel Hill, NC: Horizon Research, Inc. http://www.horizon-research.com/publications/stock.pdf

Campbell, Donald T. and Julian C. Stanley. (1966). *Experimental and Quasi-Experimental Designs for Research.* Boston, MA: Houghton Mifflin.

Cohen, Jacob. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Denzin, Norman K., and Yvonna S. Lincoln, eds. (2000). *Handbook of Qualitative Research* (2nd Ed.). Thousand Oaks, CA: Sage.

Joint Committee on Standards for Educational Evaluation. (1994). *The Program Evaluation Standards* (2nd Ed.). Thousand Oaks, CA: Sage.

National Science Foundation. (1993). *User-Friendly Handbook for Project Evaluation: Science, Mathematics, Engineering and Technology Education.* NSF 93-152. Washington, DC: NSF.

National Science Foundation. (1997). *User-Friendly Handbook for Mixed Method Evaluations.* NSF 97-153. Washington, DC: NSF.

Trochim, William M. *The Research Methods Knowledge Base.* http://trochim.human.cornell.edu/kb/index.htm